

Supervisor: Professor Gil McVean

I can supervise up to three projects in any of the areas described below. Exact details of the project will depend on the interests and background of the student. Where papers referred to below are not yet published I can give any interested student a preprint.

Methods: applied probability, computational statistics

Coalescent modelling

Patterns of genetic variation within a species are informative about diverse processes, from the action of natural selection to the fine-scale structure of recombination rate variation. In the last 15 years, coalescent models have become central to the analysis of such data^{1,2}. And recent developments in computational and statistical methods have enabled such models to be applied on a genome-wide scale^{3,4}. The project will explore mathematical⁵ and computational⁶ aspects of coalescent modelling. An area of particular interest is how coalescent approaches might be used to infer genealogical history from whole-genome data in over 10,000 individuals.

Prerequisites: BS3a Applied Probability

Statistical genetics

The use of statistical modelling to help understand the genetic basis of heritable disease has changed dramatically in the last few years. This has been driven in part by the massive increase in the scale of data available made possible by large-scale collaborations, such as the International HapMap Project⁷. For example an experiment to identify genes associated with a disease might compare genetic variation at nearly a million positions in the genome between 2,000 patients and 2,000 people without the disease⁸. It has also been driven by developments in modelling genetic variation^{4,9,10} and computational statistics. The project will explore how statistical modelling techniques are being used in large-scale disease genetics studies. Of particular focus will be the recent development of methods for predicting genetic variation at parts of the genome not directly observed in the experiment (so-called 'missing-data' problems).

Prerequisites: BS2a Foundations of Statistical Inference

Evolutionary genetics of pathogen genomes

For many human pathogens genetic variation plays an important role in influencing the interaction with its host. For example, the principle agent of malaria, *Plasmodium falciparum*, is genetically very diverse^{11,12}, a consequence of which is that obtaining successful vaccines is extremely challenging. Genome-sequencing^{13,14} and related projects¹⁵ have given some insight into just how diverse particular gene families involved in evasion of the host immune system in *P. falciparum* are. One of the great challenges in making sense of the genome data is trying to infer the influence of different evolutionary forces acting on these gene families (recombination, mutation, natural selection) from the sequences alone. Effectively, we wish to reconstruct the history of the gene sequences, but this is a difficult problem, even in the absence of recombination. This project will look at how to model the evolution of gene families involved in immune-evasion and subsequently how the evolutionary history might be inferred from gene sequences¹⁶.

Prerequisites: BS3a Applied Probability

Reference List

1. Hudson, R. R. Gene genealogies and the coalescent process in *Oxford Surveys in Evolutionary Biology* 1-44. Oxford University Press, Oxford (1990).
2. Griffiths, R. C. Ancestral inference from gene trees in *Genes, Fossils and Behaviour: An Integrated Approach to Human Evolution*, 310, 137-172. IOS Press, Amsterdam (2001).
3. McVean, G. A. & Cardin, N. J. Approximating the coalescent with recombination. *Philos.Trans.R.Soc.Lond B Biol.Sci.* 360, 1387-1393 (2005).
4. McVean, G. A. et al. The fine-scale structure of recombination rate variation in the human genome. *Science* 304, 581-584 (2004).
5. Nordborg, M. Coalescent theory in *Handbook of Statistical Genetics* 602-635. Wiley, Chichester (2003).
6. Stephens, M. Inference under the coalescent in *Handbook of Statistical Genetics*. Wiley, Chichester (2003).
7. The International HapMap Consortium. A haplotype map of the human genome. *Nature* 437, 1299-1320 (2005).
8. The Wellcome Trust Case Control Consortium. Genome-wide association studies for seven common human diseases. *Nature* 447;661-78. (2007).
9. Marchini, J., Howie, B., Myers, S., McVean, G., & Donnelly, P. A new multi-locus method for imputing genotypes in genome-wide association studies. *Nature Genetics* 39 : 906-913 (2007)
10. Li, N. & Stephens, M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* 165, 2213-2233 (2003).
11. Mu, J. et al. Genome-wide variation and identification of vaccine targets in the *Plasmodium falciparum* genome. *Nat.Genet.* 39, 126-130 (2007).
12. Mu, J. et al. Recombination hotspots and population structure in *Plasmodium falciparum*. *PLoS Biol.* 3, e335 (2005).
13. Gardner, M. J. et al. Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* 419, 498-511 (2002).
14. Jeffares, D. C. et al. Genome variation and evolution of the malaria parasite *Plasmodium falciparum*. *Nat.Genet.* 39, 120-125 (2007).
15. Barry, A. E. et al. Population Genomics of the Immune Evasion (var) Genes of *Plasmodium falciparum*. *PLoS Pathog.* 3, e34 (2007).
16. Chase, E. University of Oxford (2007).